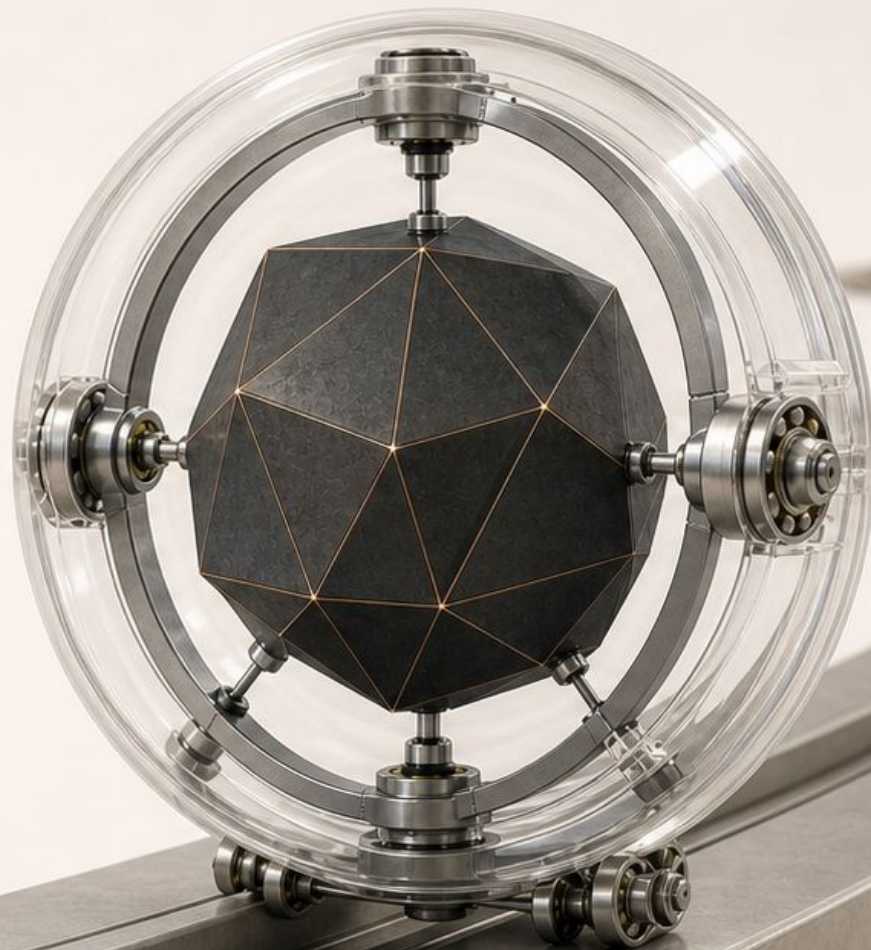


FOUNDER FORUM

Anatomy of an Agent Harness

How models become systems that can act—and keep improving.

Brennan McEachran



WHY THIS TALK

The model is only part of the product.

Two years ago, we unpacked prediction. This time, we're opening the machinery around it.

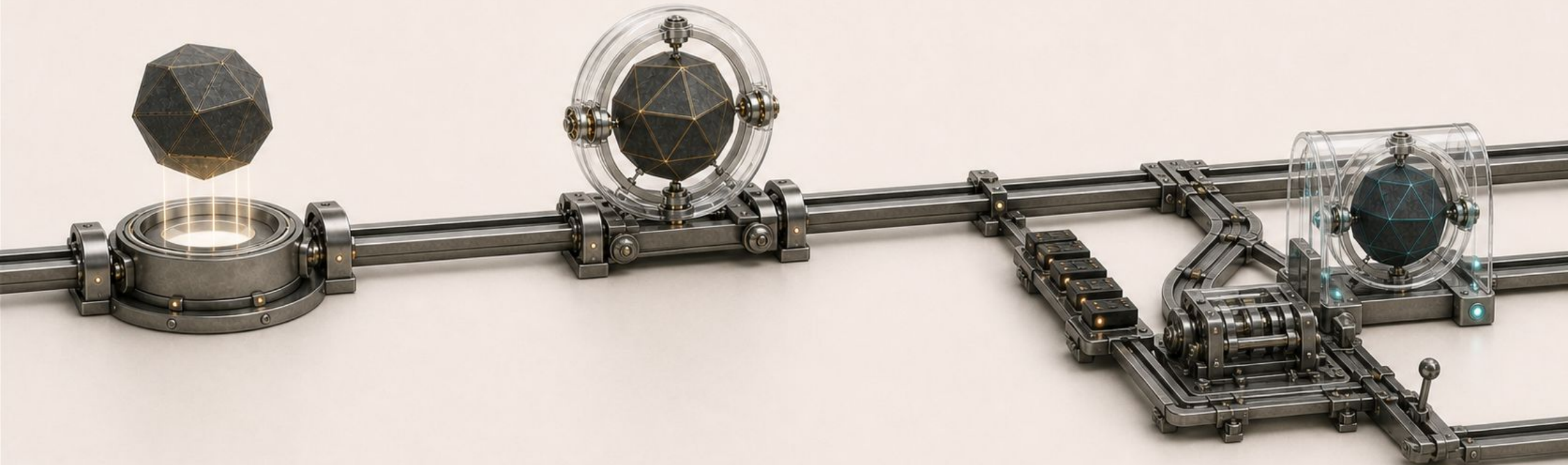
BY THE END

Biopsy an agent.
Know which part to improve.

01 THE MODEL LEARNS TO ACT

02 THE HARNESS MAKES IT A PRODUCT

03 EVIDENCE IMPROVES THE SYSTEM

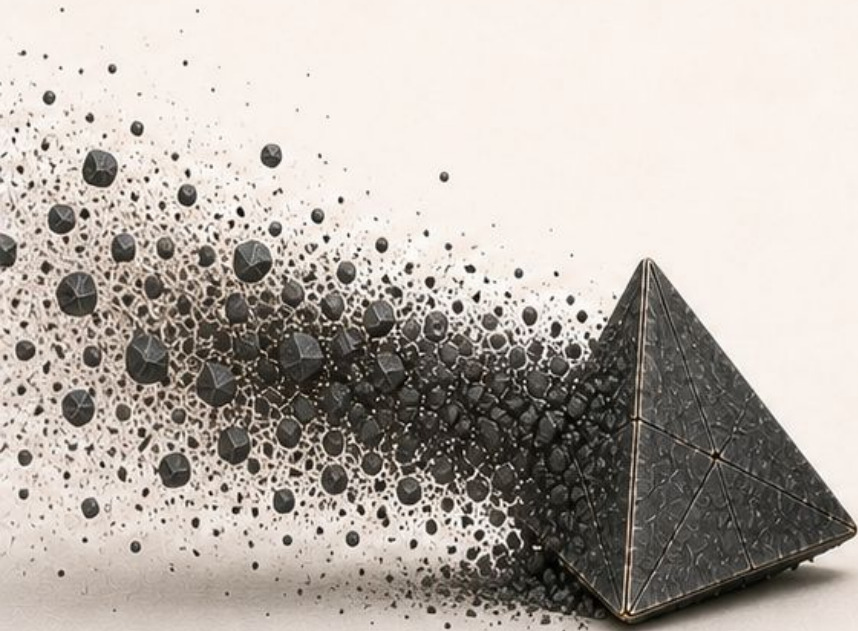


PART I

How the model learns to act

Autocomplete → assistant → tools → persistent behavior

TRAINING REFINES THE SAME CORE



BASE MODEL



ASSISTANT

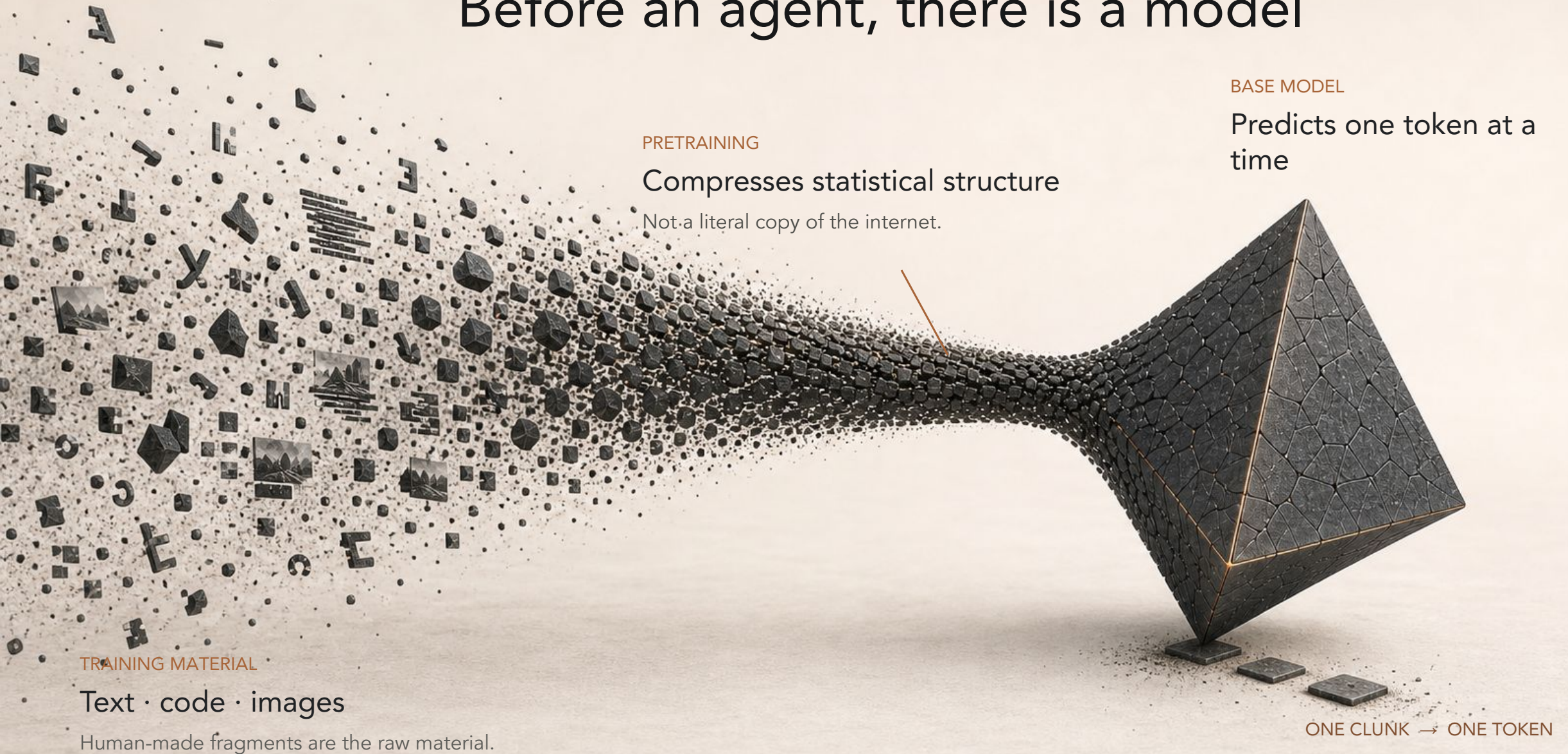


TOOL USER



AGENTIC MODEL

Before an agent, there is a model



Post-training gives the model a role

A conversational protocol says who is speaking. Training teaches that participant how to behave.

PUBLIC MESSAGE STRUCTURE

system	OpenAI policy
developer	App instructions
user	End-user request
	↓
assistant	Model output

Roles structure the conversation and its authority.

ROLE PROTOCOL

BEHAVIORAL TRAINING

UNDERNEATH

It still predicts the next token.



Tool training gives the model new keys

The model proposes. Software validates. The calculator acts.

REPRESENTATIVE TRACE

USER

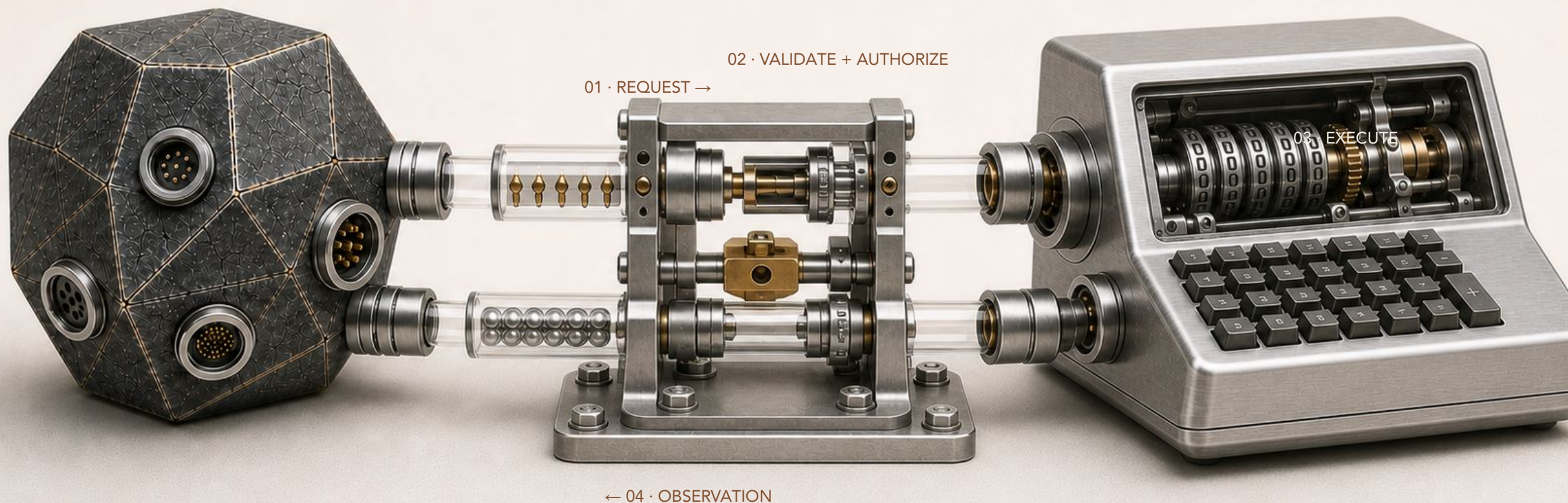
What is 317×89 ?

MODEL → TOOL

{ "expression": "317*89" }

TOOL → MODEL

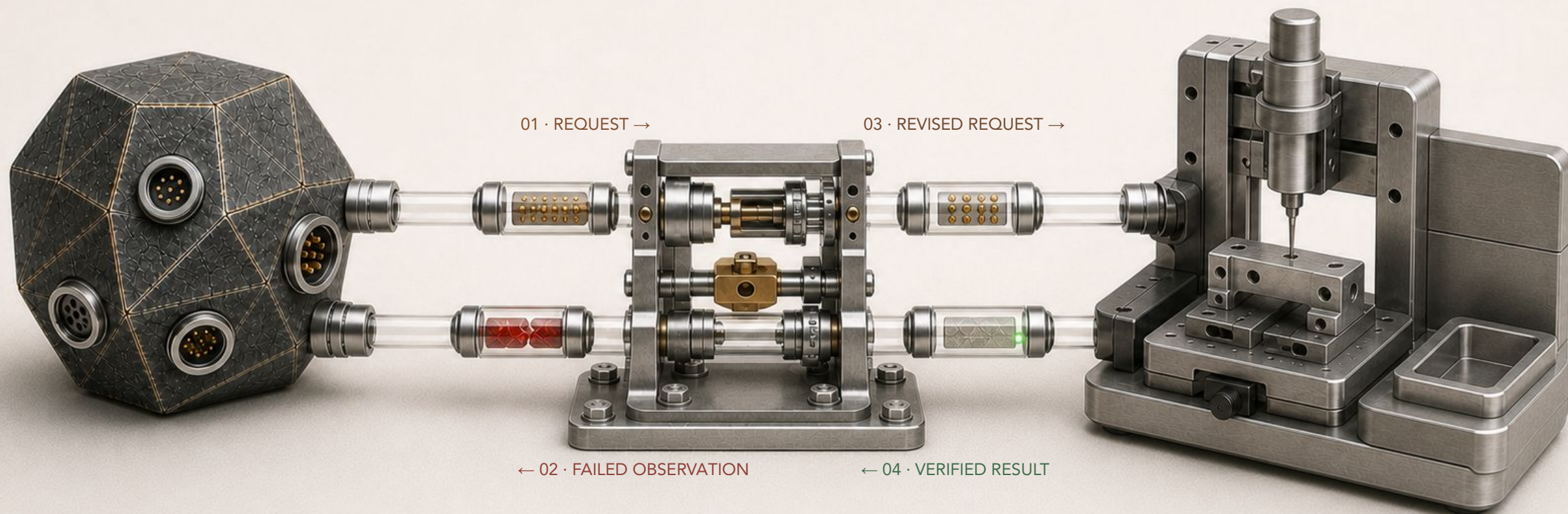
28213



Agentic models pursue outcomes, not answers

The same fixed model can use a failed observation to change its next action.

More inference-time work
can improve performance.



PART II

The model can act. It still isn't a product.

The loop supplies persistence. The harness makes that capability operable.

MODEL
Capability

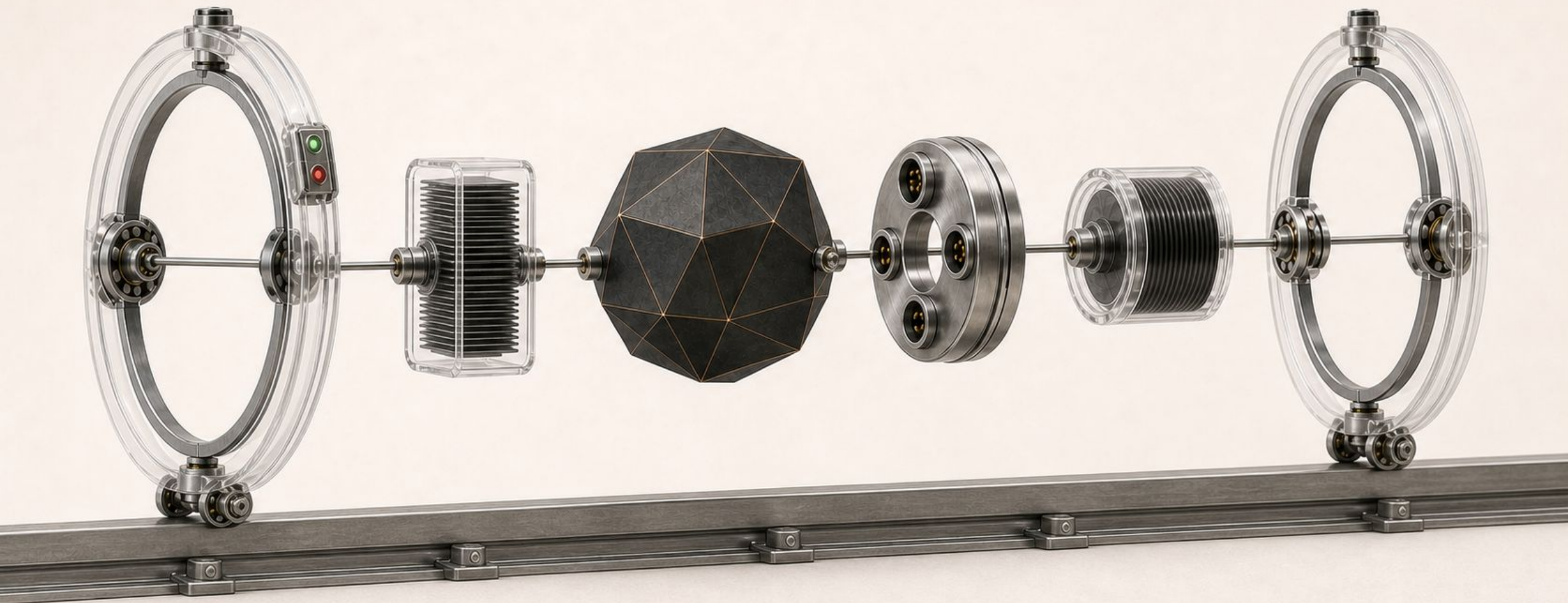
LOOP
Persistence

HARNESS
Context · control
continuity · evidence



Biopsy of an agent harness

One rolling object. Five separable responsibilities.



GUARDRAILS

SYSTEM PROMPT

MODEL

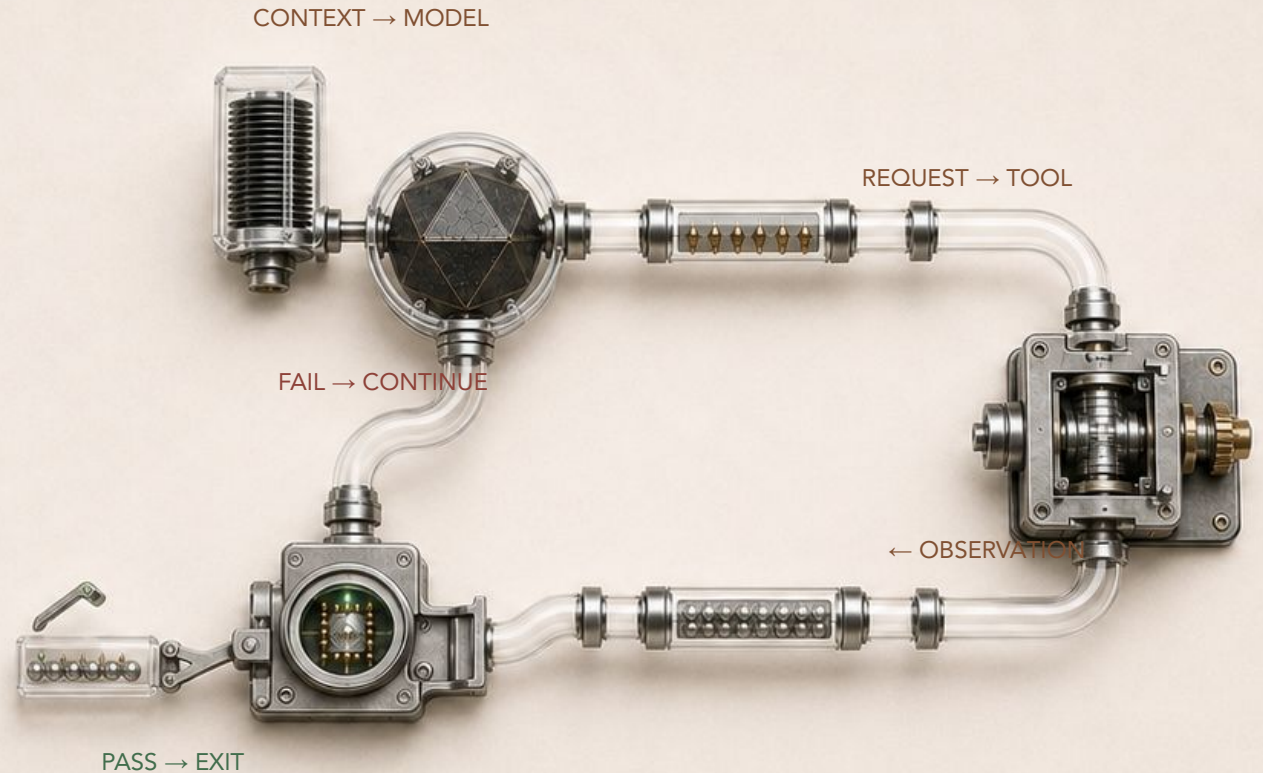
TOOLS

MEMORY

The loop at the center fits on a napkin

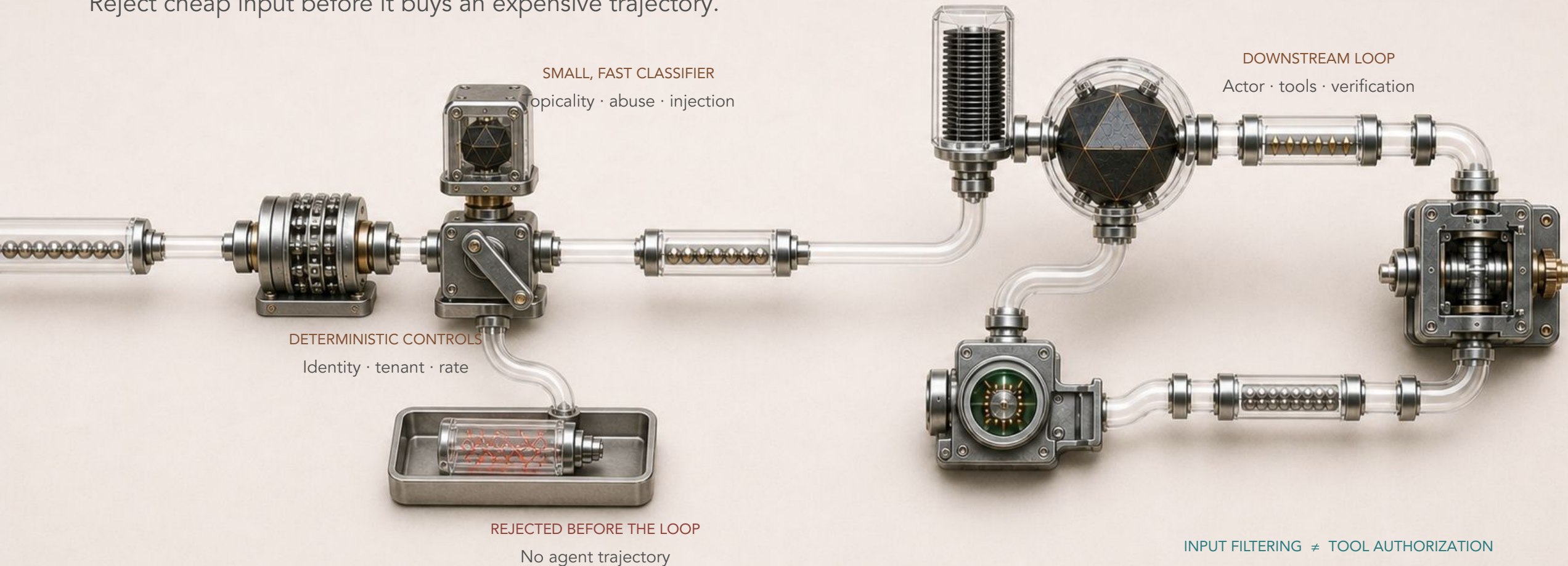
The harness keeps extending one conversation.

```
01 while (withinBudget) {  
02   step = model(context)  
03   if (step.tool) context += execute(step.tool)  
04   else if (verify(step.answer)) return step.answer  
05   else context += failureEvidence  
06 }
```



Guard the expensive loop before it starts

Reject cheap input before it buys an expensive trajectory.



The conversation begins before the user speaks

The user message is the final layer—not the whole prompt.

STABLE PREFIX

Policy · developer · tools / skills

DYNAMIC TAIL

History · memory · runtime state

ASSEMBLED CONTEXT

One serialized request

MODEL CALL

Same fixed model



USER MESSAGE · INSERTED LAST

Stable prefixes preserve cacheability.

A tool is a contract attached to an executor

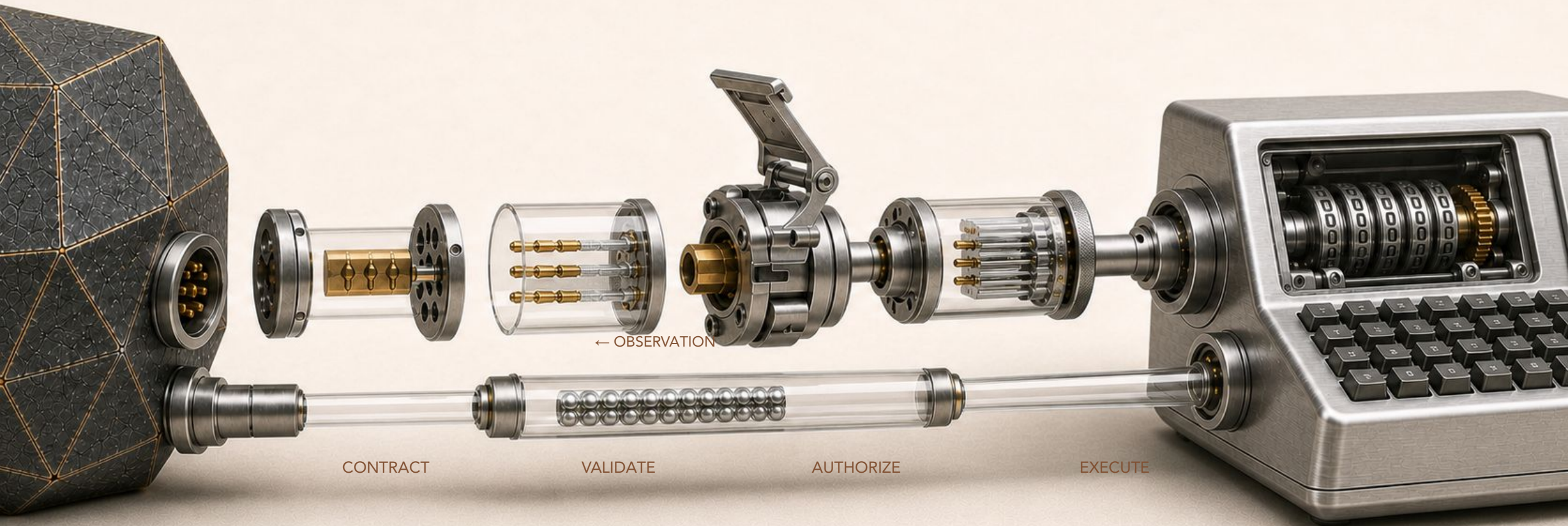
The model proposes. The harness controls what can actually happen.

MODEL-VISIBLE CONTRACT

calculator

expression: string

→ number | error



Skills load procedures only when needed

Catalog metadata stays visible. The selected playbook enters context.

REAL CATALOG ENTRY

Presentations

Create or edit PowerPoint
or Google Slides decks

CATALOG

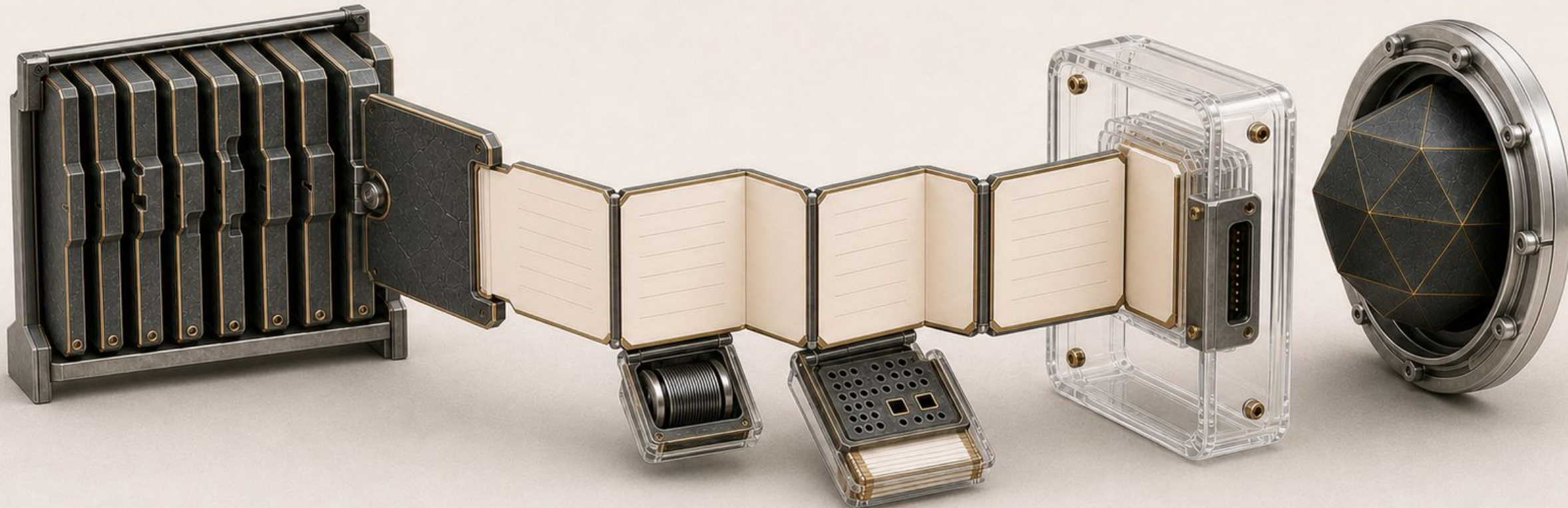
Name + one-line description

SELECTED SKILL

SKILL.md expands
+ scripts · templates · references

LOADED CONTEXT

Token cost begins here

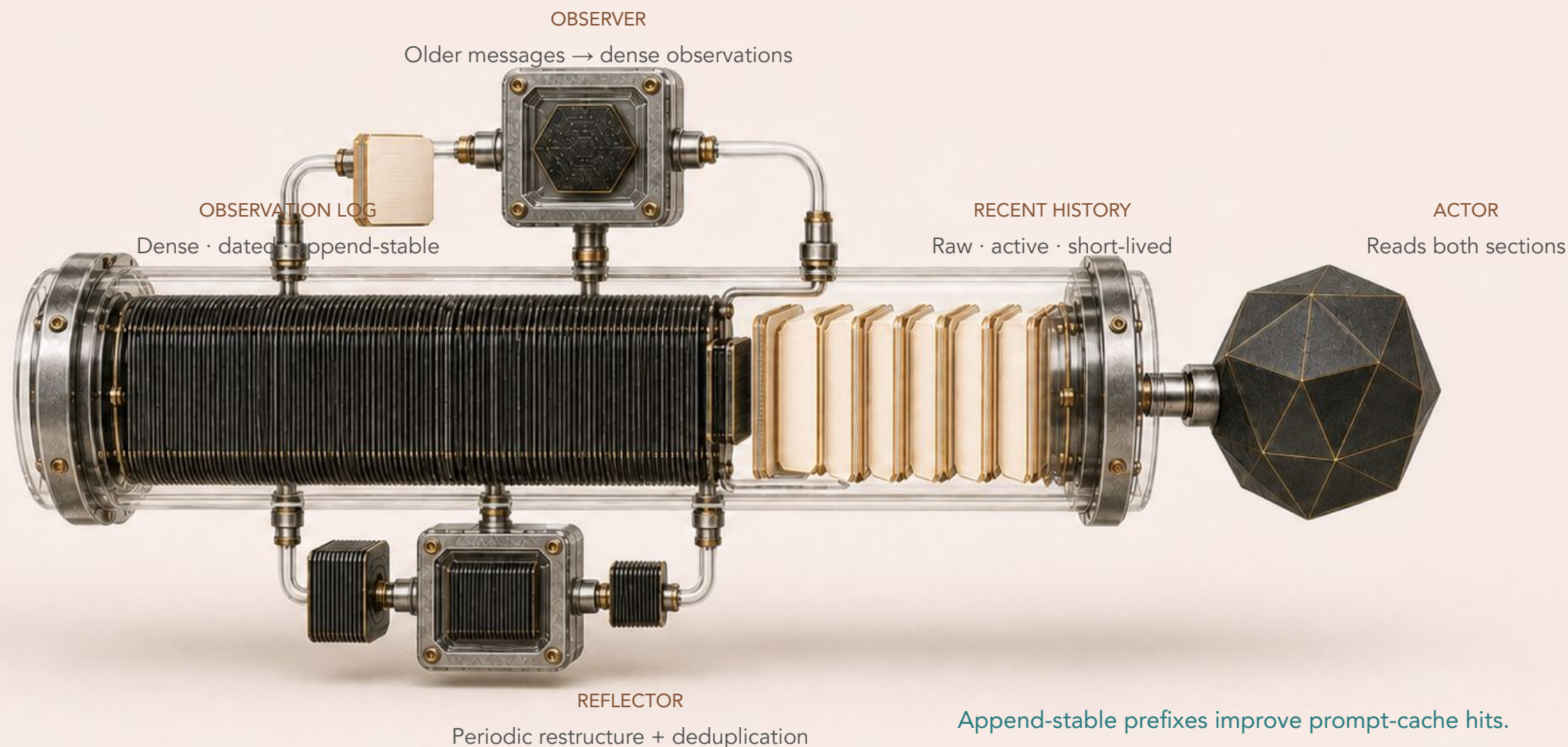


Long-running agents need selective memory

The actor sees a stable log plus recent history.

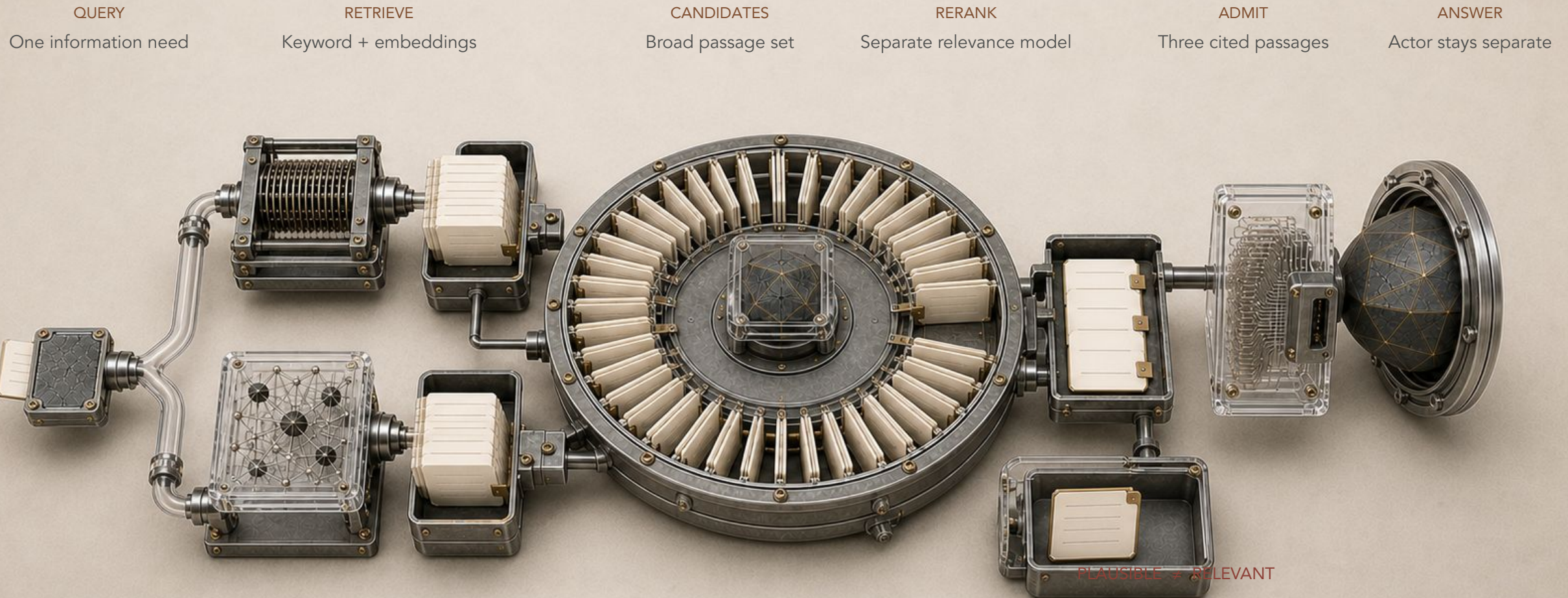
94.87%

LongMemEval · GPT-5 mini actor
Observer / Reflector: Gemini 2.5 Flash



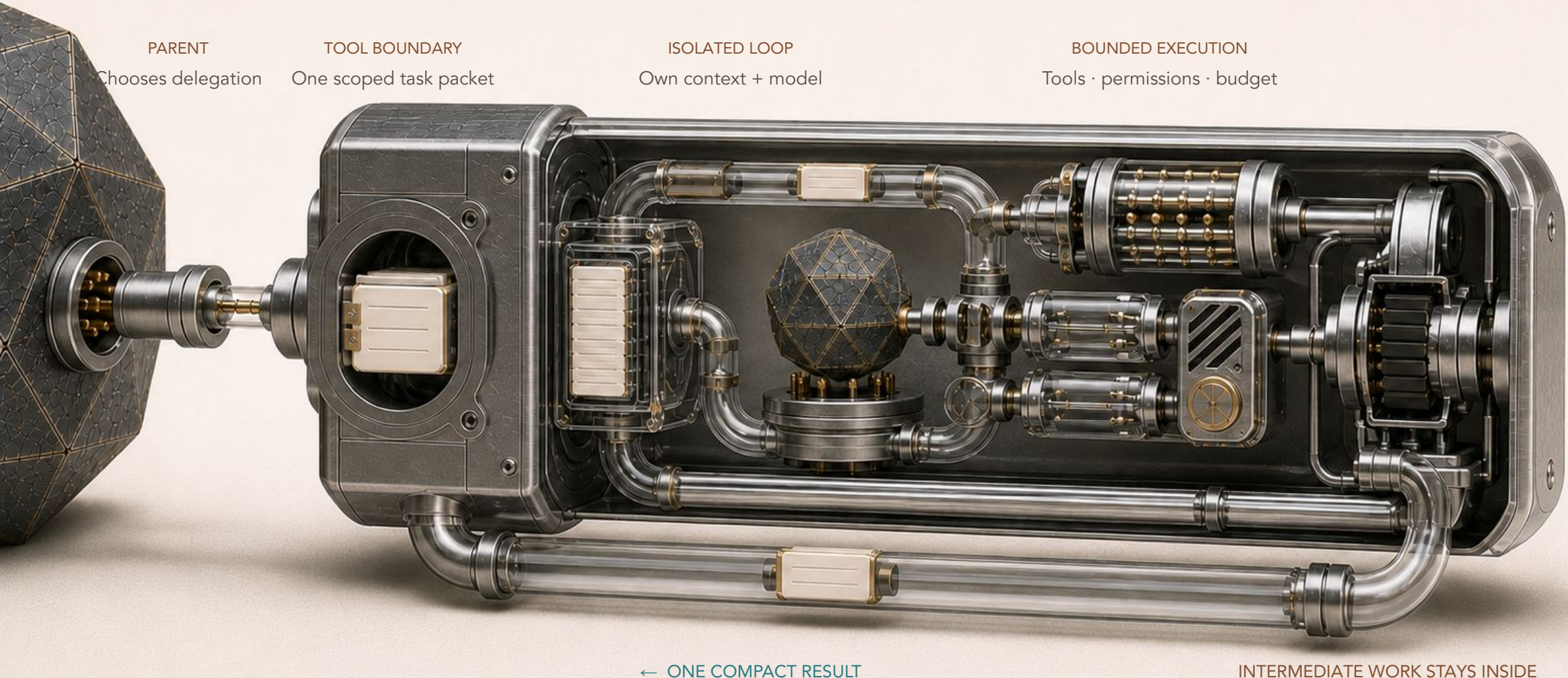
Ranking decides what earns context

RAG is a search tool containing retrieval, ranking, and evidence packaging.



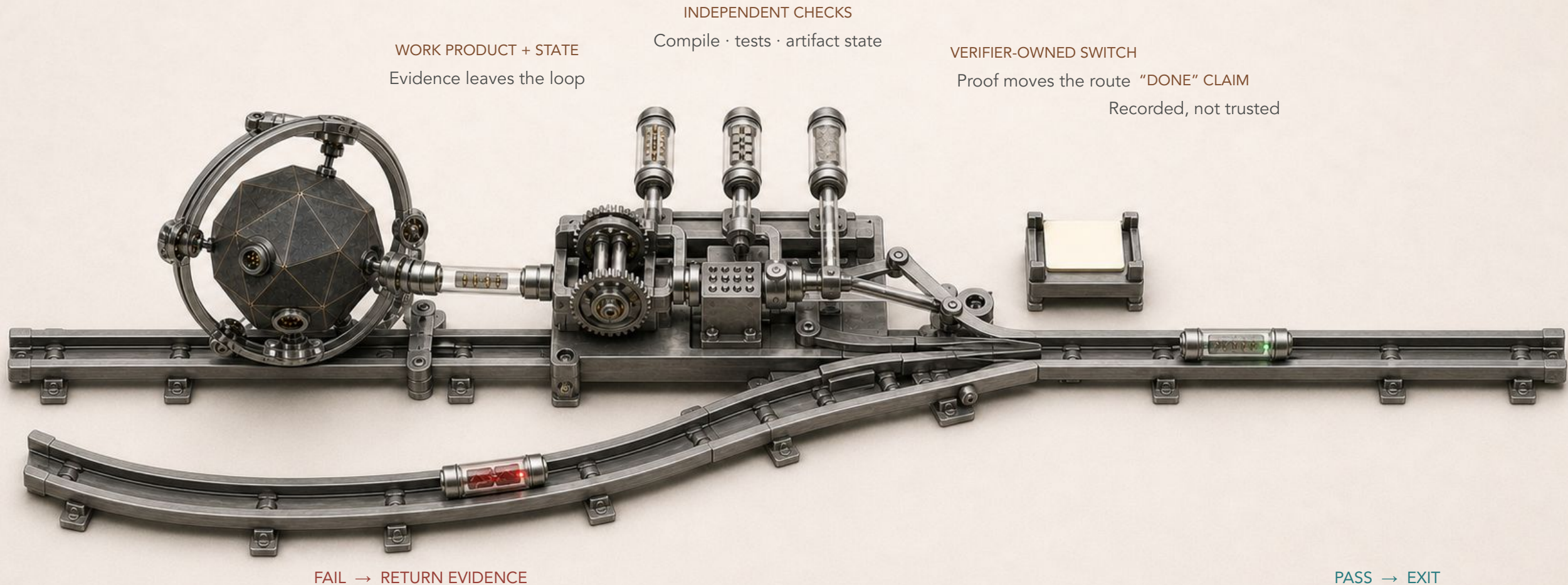
A subagent is another agent behind a tool call

The parent sends scope—not its whole context—and gets one result back.



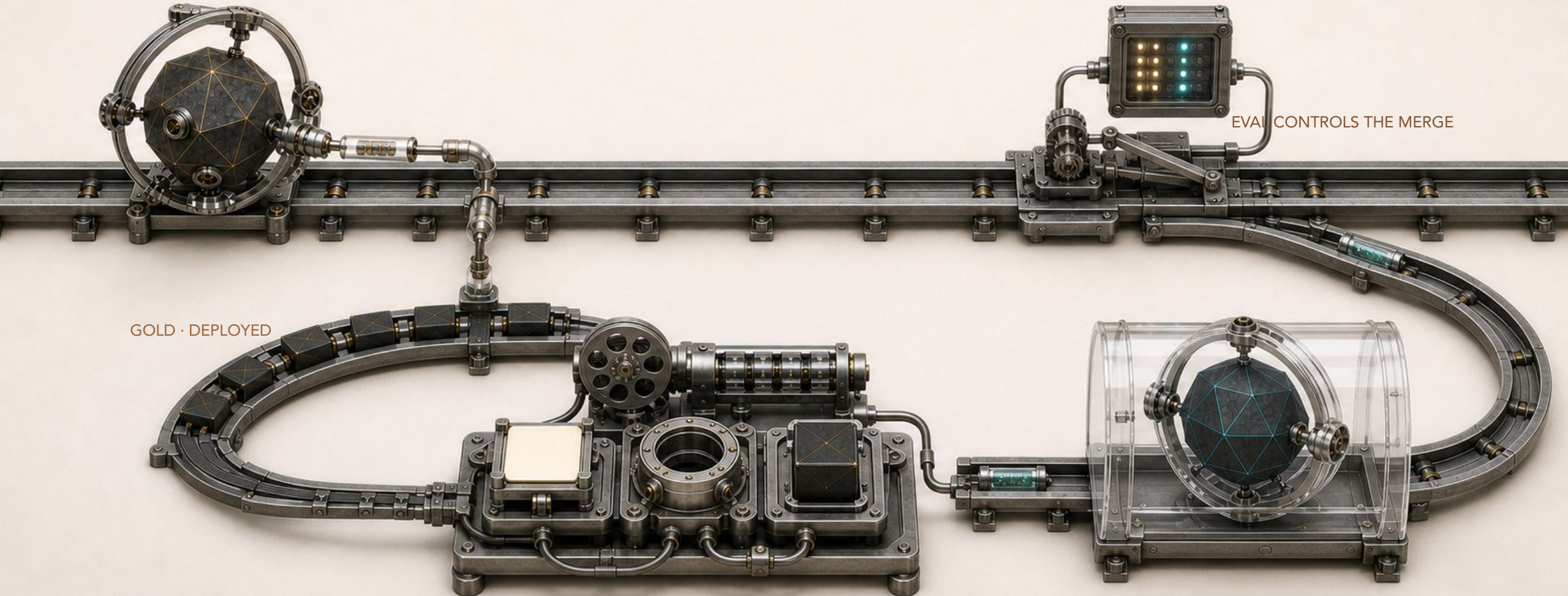
Confidence is not an exit condition

The model proposes completion. Independent evidence decides whether the loop exits.



How evidence improves the system

Measure every layer. Replay failures. Promote only what wins.



GOLD · DEPLOYED

EVAL CONTROLS THE MERGE

EVIDENCE → REPLAY → EXPERIMENT

TEAL · CANDIDATE · QUARANTINED

There is no single model behind the agent

Each job wants a different mix of intelligence, context, latency, modality, and price.

GUARD

Fast + cheap

MEMORY

Long context

EMBED

Vectors at scale

RERANK

Relevance

VERIFY

Independent judge

IMAGE

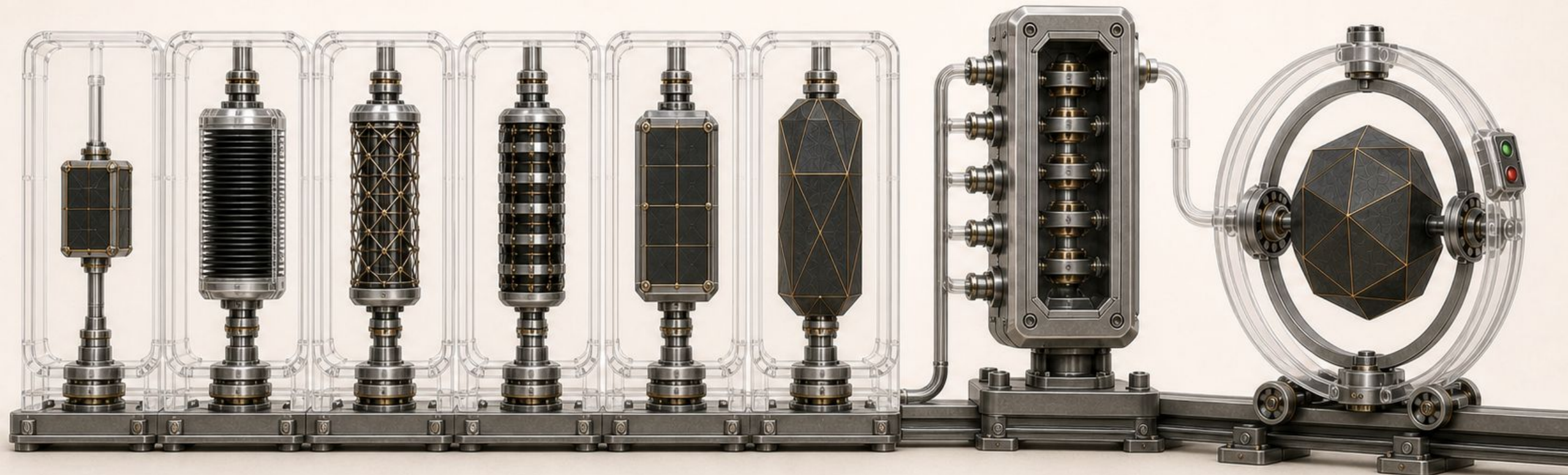
Multimodal

GATEWAY

Access + accounting

ACTOR

Plan + act



GATEWAYS STANDARDIZE ACCESS. EVALS MAKE REPLACEMENT SAFE.

Public benchmarks scout. Private evals decide.

AA-Omniscience screens for knowledge reliability and hallucination—not your product.

PUBLIC MARKET

450

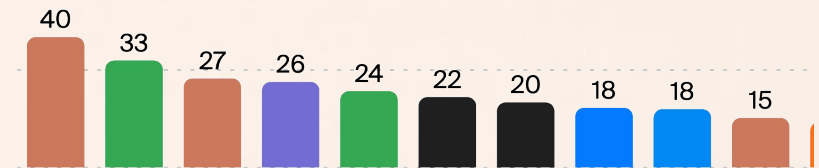
models listed

Filter by the job: reliability · context ·
latency · cost · modality

JOB-CONSTRAINED SHORTLIST

3–5

worth testing privately



AI
Claude Fable 5 (with fallback)

G
Gemini 3.1 Pro Preview

AI
Claude Opus 4.8 (max)

X
Grok 4.5 (high)

G
Gemini 3.6 Flash

🌀
GPT-5.6 Sol (max)

🌀
GPT-5.5 (xhigh)

K
Kimi K3

∞
Muse Spark 1.1 (xhigh)

AI
Claude Sonnet 5 (max)

Qwen3.7

Be loyal to the eval, not the model.

Same private cases. Same rubric. Same settings. Then plot quality against cost.

VERSIONED PRIVATE EVAL

The job becomes cases + a rubric

SAME CASES · SAME RUBRIC · SAME BUDGET

PUBLIC SCOUTING VIEW · INTELLIGENCE × COST

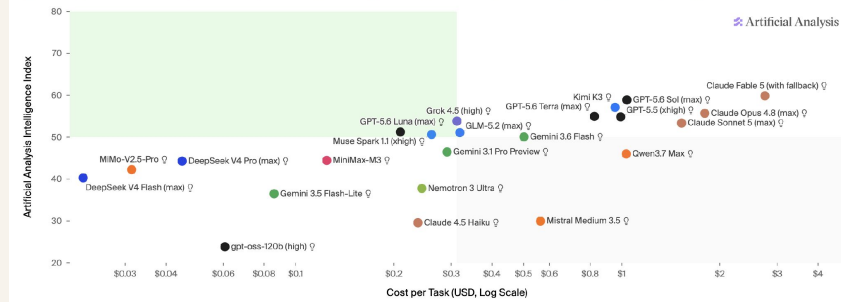
Intelligence Index vs. Cost per Intelligence Index Task

Artificial Analysis Intelligence Index · Weighted average cost (USD) per Artificial Analysis Intelligence Index task

28 of 581 models

Most attractive quadrant

● Xiaomi ● Google ● OpenAI ● Anthropic ● MiniMax ● NVIDIA ● Alibaba ● SpaceXAI ● DeepSeek ● Meta ● Mistral ● Z AI ● Kimi



THE SHAPE IS PUBLIC. YOUR FRONTIER MUST COME FROM YOUR CASES.

PRODUCT PRIORITY CHOOSES THE COMPROMISE

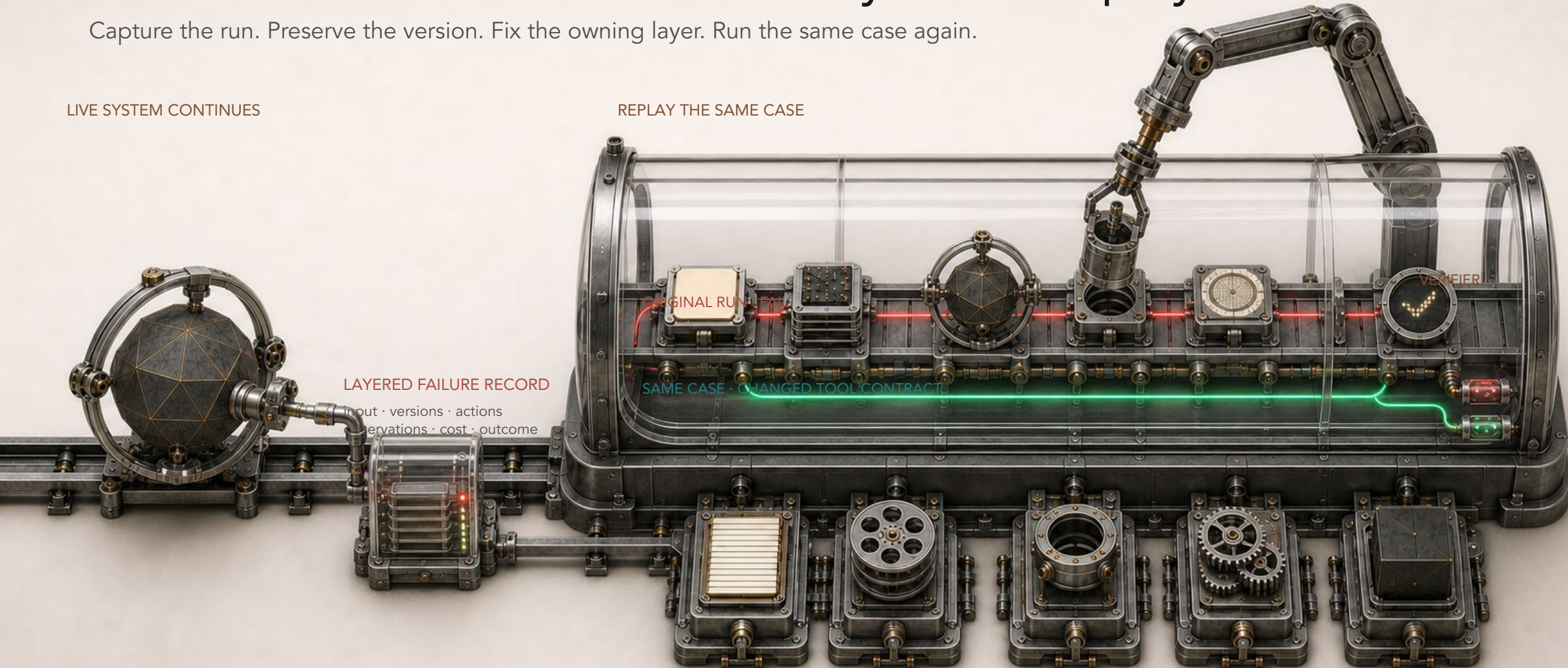
quality · cost · latency

A failure becomes an asset when you can replay it.

Capture the run. Preserve the version. Fix the owning layer. Run the same case again.

LIVE SYSTEM CONTINUES

REPLAY THE SAME CASE



LAYERED FAILURE RECORD

input · versions · actions
observations · cost · outcome

ORIGINAL RUN - FAILURE

SAME CASE · CHANGED TOOL CONTRACT

VERIFIER

FIRST DESTINATION: A VERSIONED REGRESSION TEST

Not model training—yet.

PROMPT /
CONTEXT

RETRIEVAL /
DATA

TOOL
CONTRACT

WORKFLOW /
PERMISSIONS

MODEL

Fine-tuning is the last mile, not the first move.

Train only when the behavior is demonstrable, stable, measurable—and still missing.

INCUMBENT STAYS LIVE

STARTING MODEL COPY + CURATED CASES + OBJECTIVE

A controlled weight-update experiment

MOST FAILURES: FIX THE HARNESS

FEW SELECTED CASES

SAME EVAL CONTROLS PROMOTION



CONTEXT

RETRIEVAL

TOOLS

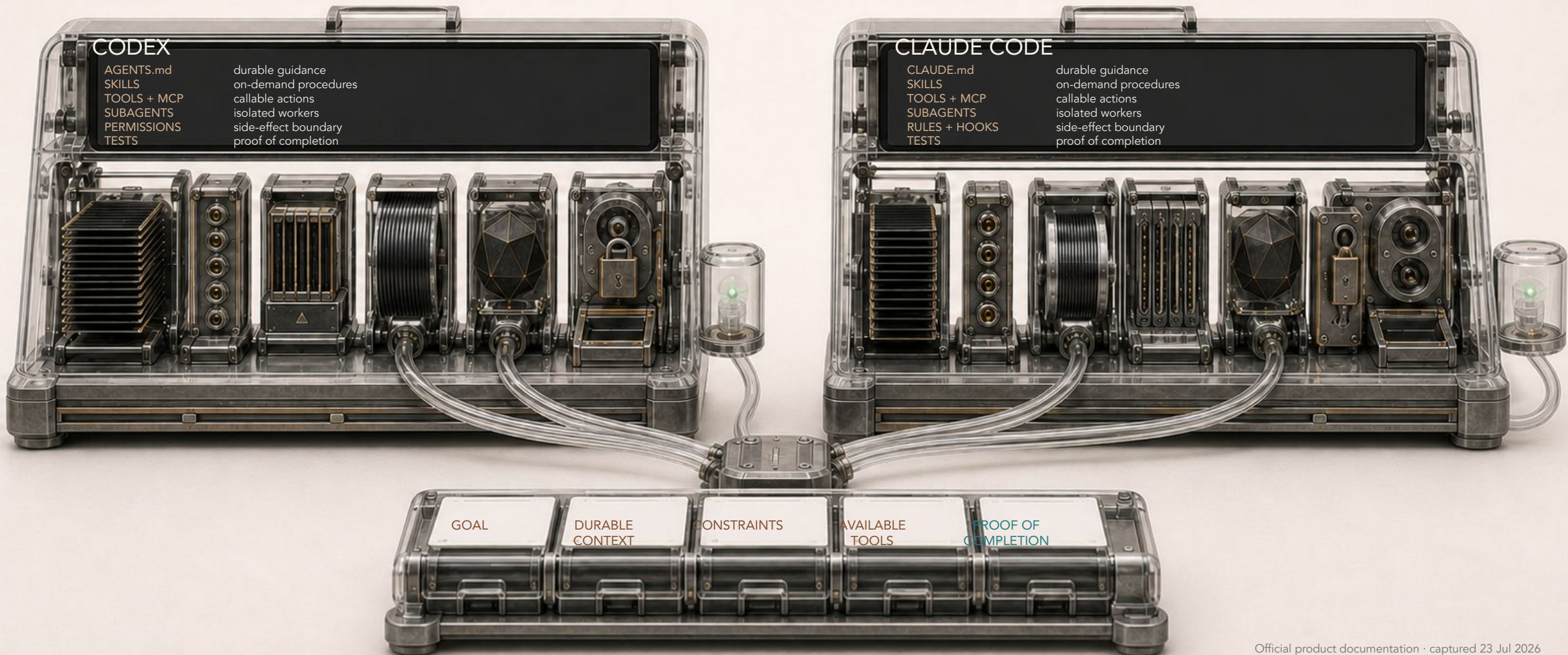
WORKFLOW /
PERMISSIONS

TEAL = NEW CANDIDATE WEIGHTS

TURNOUT CLOSED

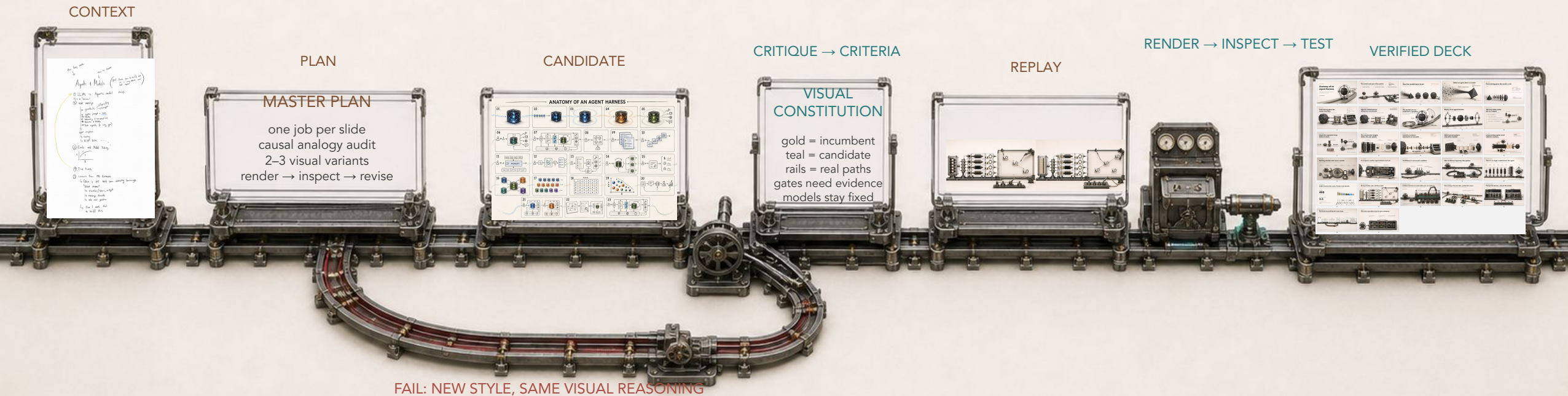
Prompt the harness, not just the model.

Codex and Claude Code expose corresponding responsibilities through different implementations.



This deck was built by the same loop.

Sketches became a plan. Rejected outputs became criteria. Every slide was rendered, inspected, and revised.



Every object on this rail is a real artifact from this task.

TAKE IT BACK

Take three questions back to your company.

If you can answer them clearly, the rest of the harness becomes concrete.

01 WHAT IS THE LOOP?

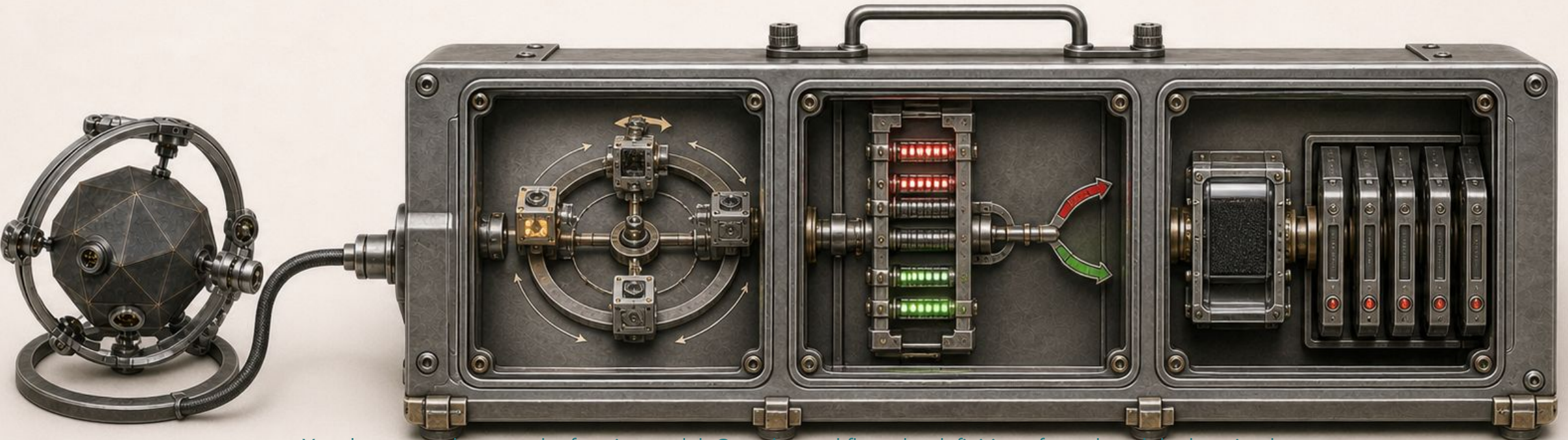
What can the agent observe, decide, and do?

02 WHAT DEFINES DONE?

Which independent evidence opens the exit?

03 WHICH FAILURES DO WE KEEP?

What becomes a permanent eval?



You do not need to own the frontier model. Own the workflow, the definition of good, and the learning loop.